Nucleic Acids Research, 2004, Vol. 32, Database issue ***D523–D527***
DOI: 10.1093/nar/gkh013

RTCGD: retroviral tagged cancer gene database

Keiko Akagi, Takeshi Suzuki, Robert M. Stephens[1], Nancy A. Jenkins and Neal G. Copeland*

Mouse Cancer Genetics Program, National Cancer Institute, Frederick, MD 21701, USA and [1]ABCC Science Applications International Corporation, Frederick, MD 21702, USA

Received July 28, 2003; Revised August 14, 2003; Accepted September 3, 2003

ABSTRACT

Retroviral insertional mutagenesis in mouse hematopoietic tumors provides a potent cancer gene discovery tool in the post-genome-sequence era. To manage multiple high-throughput insertional mutagenesis screening projects, we developed the Retroviral Tagged Cancer Gene Database (RTCGD; http://RTCGD.ncifcrf.gov). A sequence analysis pipeline determines the genomic position of each retroviral integration site cloned from a mouse tumor, the distance between it and the nearest candidate disease gene(s) and its orientation with respect to the candidate gene(s). The pipeline also identifies genomic regions that are targets of retroviral integration in more than one tumor (common integration sites, CISs) and are thus likely to encode a disease gene. Users can search the database using a specified gene symbol, chromosome number or tumor model to identify both CIS genes and unique viral integration sites or compare the integration sites cloned by different laboratories using different models. As a default setting, users first review the CIS Lists and then Clone Lists. CIS Lists describe CISs and their candidate disease genes along with links to other public databases and clone lists. Clone Lists describe the viral integration site clones along with the tumor model and tumor type from which they were cloned, candidate disease gene(s), genomic position and orientation of the integrated provirus with respect to the candidate gene(s). It also provides a pictorial view of the genomic location of each integration site relative to neighboring genes and markers. Researchers can identify integrations of interest and compare their results with those for multiple tumor models and tumor types using RTCGD.

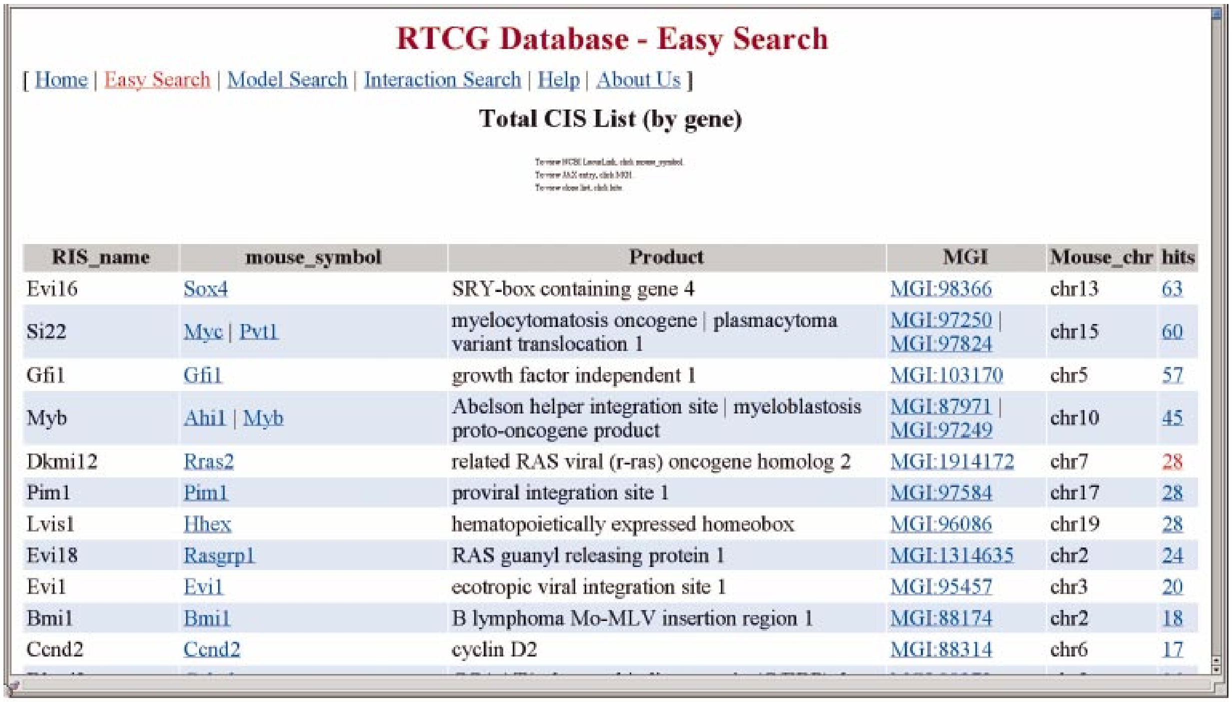
INTRODUCTION

Retroviruses induce hematopoietic tumors in mice by integrating into the genome and deregulating the expression of proto-oncogenes or inactivating the expression of tumor suppressor genes [reviewed in (1)]. The retroviral integration sites (RISs) in the tumors thus provide powerful molecular tags for the discovery of genes involved in cancer. During the past few years, two advances have dramatically increased our ability to identify cancer genes via insertional mutagenesis. The first advance was the development of high-throughput PCR-based methods for cloning retroviral integration sites along with mouse flanking cellular DNA from tumors (2–5). The second advance was the publication of the mouse genome sequence (6). Now, mouse cellular DNA sequences flanking retroviral integration sites in tumors can be BLAT (7) searched against the nearly complete mouse genome sequence and candidate disease genes rapidly evaluated.

Recently, several high-throughput insertional mutagenesis screens were published (1,8–11). These screens led to the identification of 236 common integration sites (CISs). For 93.6% (221/236) of these CISs, candidate disease genes could be identified simply by examining the annotated mouse genome sequence. Several of these genes are validated human cancer genes, while many others have not yet been examined for a role in cancer. These are the most interesting since they represent potentially new genes involved in cancer. In many tumors more than one disease gene was identified by insertional mutagenesis. This was not unexpected given that most tumors contain multiple retroviral integrations and cancer is known to result from the accumulation of multiple mutations that cooperate to induce disease. When two or more disease genes are identified in the same tumor by insertional mutagenesis this result provides strong genetic evidence for cooperativity between these genes in disease induction. This kind of genetic evidence has led to the identification of a new class of cofactors for HOX and PBX homeodomain proteins (i.e. the MEIS proteins) and helped to explain the functional diversity among homeodomain proteins (12). These studies demonstrate the power of retroviral insertional mutagenesis for cancer gene discovery in the post-genome-sequence era.

Comparison of insertional mutagenesis data generated in different laboratories has identified a set of CIS genes that are unique to each screen, while other CIS genes appear more common and are found in two or more screens. In contrast, some CIS genes are very promiscuous and were identified in all or most screens (1,9). Genes identified in multiple screens using different mouse models have an increased probability of representing true disease genes. These results, combined with

*To whom correspondence should be addressed. Tel: +1 301 846 1260; Fax: +1 301 846 6666; Email: Copeland@ncifcrf.gov
Present address:
T.Suzuki, Cancer Genetics Unit, Horizontal Medical Research Organization, Kyoto University, Kyoto 606-8501, Japan



RTCG Database - Easy Search

[Home | Easy Search | Model Search | Interaction Search | Help | About Us]

Total CIS List (by gene)

To view NCBI LocusLink, click mouse_symbol.
To view MGI entry, click MGI.
To view clone list, click hits.

RIS_name	mouse_symbol	Product	MGI	Mouse_chr	hits
Evi16	Sox4	SRY-box containing gene 4	MGI:98366	chr13	63
Si22	Myc | Pvt1	myelocytomatosis oncogene | plasmacytoma variant translocation 1	MGI:97250 | MGI:97824	chr15	60
Gfi1	Gfi1	growth factor independent 1	MGI:103170	chr5	57
Myb	Ahi1 | Myb	Abelson helper integration site | myeloblastosis proto-oncogene product	MGI:87971 | MGI:97249	chr10	45
Dkmi12	Rras2	related RAS viral (r-ras) oncogene homolog 2	MGI:1914172	chr7	28
Pim1	Pim1	proviral integration site 1	MGI:97584	chr17	28
Lvis1	Hhex	hematopoietically expressed homeobox	MGI:96086	chr19	28
Evi18	Rasgrp1	RAS guanyl releasing protein 1	MGI:1314635	chr2	24
Evi1	Evi1	ecotropic viral integration site 1	MGI:95457	chr3	20
Bmi1	Bmi1	B lymphoma Mo-MLV insertion region 1	MGI:88174	chr2	18
Ccnd2	Ccnd2	cyclin D2	MGI:88314	chr6	17

Figure 1. A typical CIS List from RTCGD. Each row represents a CIS. For each CIS the retroviral integration site name (RIS name), candidate gene (by official mouse symbol), gene product description, Mouse Genome Database ID (MGI), mouse chromosomal location and number of independent retroviral integrations at the CIS with a link to the corresponding Clone List are given.

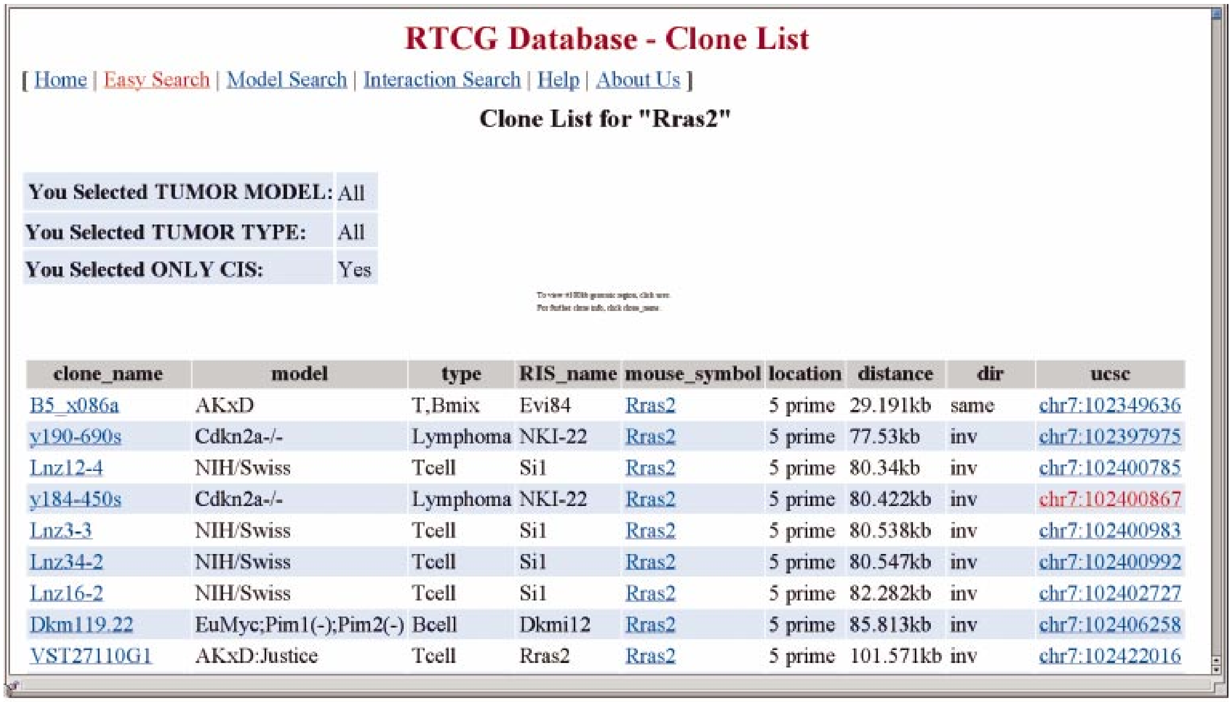
the fact that insertional mutagenesis data are cumulative and hundreds of candidate disease genes are likely to be identified by insertional mutagenesis in the future has prompted us to develop the Retroviral Tagged Cancer Gene Database (RTCGD). RTCGD provides a web-based front end, which allows researchers to identify published CIS genes and corresponding retroviral integration site sequences for any chromosome or mouse genomic region of interest. It also enables users to compare CISs identified in different laboratories using different mouse models and identify the CIS genes that are the most promiscuous. Finally, RTCGD makes it possible to easily identify CIS genes that are mutated in the same tumor and hence likely to cooperate to induce disease. We believe that RTCGD will become a widely used resource for researchers studying cancer using mouse models and/or human tumors.

RTCGD CONTENT

At the time of submission, the RTCGD contained 3100 retroviral integration site sequences cloned from 17 mouse tumor models by nine different laboratories. These 3100 integration sites define 236 CISs. A list of these CISs as well as unique integration sites is provided to the user via list pages. These list pages are generated dynamically, providing the user with the most recent information. Data fields on the CIS List pages include CIS name, mouse symbol for each CIS candidate disease gene, information regarding the function of protein expressed at each CIS along with links to the Mouse Genome Informatics (MGI) web site, the mouse chromosomal location of each CIS and the number of independent retroviral integrations at each CIS (Fig. 1). On the Clone List pages, the tumor model and tumor type from which the retroviral integration was cloned are listed. Chromosome and nucleotide position of the integration site on the mouse genome are also provided for the public version of the mouse genome sequence (6) (Fig. 2). In addition, the relative location, distance and orientation of each retroviral integration relative to the CIS candidate disease gene(s), determined using the Ensembl or NCBI server, is described for each CIS candidate gene. There are also two associated data pages. One is Clone Data, which describes the mouse sequence flanking the retroviral integration that was used to determine its position in the mouse genome and the tumor's background information. Another is Tumor Data, which lists associated viral integration clones.

RTCGD NAVIGATION

The RTCGD home page (http://RTCGD.ncifcrf.gov) provides easy access to any part of the database. Links are provided under the RTCGD banner to three search pages (Easy Search, Model Search and Interaction Search) and two static basic information pages (About Us and Help sections). These three search pages allow users to access list pages of interest and guide the user to other data through links to other web sites. Easy Search is a tool to search RTCGD based on CIS locus name, CIS candidate gene symbol or mouse chromosome number. To enhance the usability of RTCGD, we allow users to retrieve all CISs as CIS List pages by leaving a textbox blank (Fig. 1). CIS List pages display CIS names (RIS names), CIS candidate gene symbols, product description and the number of independent viral integrations for each CIS. As a



RTCG Database - Clone List

[Home | Easy Search | Model Search | Interaction Search | Help | About Us]

Clone List for "Rras2"

You Selected TUMOR MODEL:	All
You Selected TUMOR TYPE:	All
You Selected ONLY CIS:	Yes

To view ±100kb genomic region, click ucsc.
For further clone info, click clone_name.

clone_name	model	type	RIS_name	mouse_symbol	location	distance	dir	ucsc
B5_x086a	AKxD	T,Bmix	Evi84	Rras2	5 prime	29.191kb	same	chr7:102349636
y190-690s	Cdkn2a-/-	Lymphoma	NKI-22	Rras2	5 prime	77.53kb	inv	chr7:102397975
Lnz12-4	NIH/Swiss	Tcell	Si1	Rras2	5 prime	80.34kb	inv	chr7:102400785
y184-450s	Cdkn2a-/-	Lymphoma	NKI-22	Rras2	5 prime	80.422kb	inv	chr7:102400867
Lnz3-3	NIH/Swiss	Tcell	Si1	Rras2	5 prime	80.538kb	inv	chr7:102400983
Lnz34-2	NIH/Swiss	Tcell	Si1	Rras2	5 prime	80.547kb	inv	chr7:102400992
Lnz16-2	NIH/Swiss	Tcell	Si1	Rras2	5 prime	82.282kb	inv	chr7:102402727
Dkm119.22	EuMyc;Pim1(-);Pim2(-)	Bcell	Dkmi12	Rras2	5 prime	85.813kb	inv	chr7:102406258
VST27110G1	AKxD:Justice	Tcell	Rras2	Rras2	5 prime	101.571kb	inv	chr7:102422016

Figure 2. A typical Clone List from RTCGD. Clone names are linked to the corresponding Clone Data, which include the flanking mouse sequence and tumor data. Also included are tumor type, CIS name (RIS name), location of the integration site with respect to the candidate gene(s), distance from the candidate gene and orientation relative to the candidate gene. Genomic position data (ucsc) are linked to the UCSC mouse genome viewer with RTCGD custom annotation tracks.

default, the search is limited to CISs, but users can also search non-common integration sites (unique integrations). Model Search is a tool to retrieve CIS, Clone and Tumor lists for various tumor models and tumor types. As a default, the query results are displayed as CIS Lists, but users can change the setting to review Tumor Lists or Clone Lists based on their needs. Interaction Search is a tool to identify genes that might cooperate to induce disease. It displays a list of CIS genes that are also targets of retroviral integration in the same tumor containing an integration near the query gene of interest.

Each list page contains links to other information. For example, on the CIS List page, the number of integrations field for each CIS acts as a link to the corresponding Clone List (Fig. 1). On the Clone List, there are links to Clone Data, which includes the flanking sequence and tumor information (Fig. 2). Links are also available to the candidate gene on the NCBI server (13) as well as to the genomic region on the UCSC mouse genome server (14) from Clone List. We created a custom URL to display our original custom annotation tracks on the UCSC genome browser (Fig. 3), which provides an easy graphical way to determine the location and orientation of each retroviral integration with respect to flanking candidate genes (14,15).

RTCGD IMPLEMENTATION

The data model of RTCGD comprises six tables with tumors and integration sites treated as central objects. Currently, MySQL 3.23.27-beta running on a Sun Solaris is used for database management. All user access to the system is currently provided through the web interface. Pages are dynamically generated via Perl-CGI scripts, which integrate the results of queries on the database into various HTML templates. In addition to the public access pages, we have developed web editors that allow project members to enter new data not generated automatically.

We have also developed a sequence analysis pipeline to populate the data fields derived from the public genome assembly and annotation. Such a system is necessary given the increasing sophistication of genomic annotation (16) and the increasing number of clones deposited in our database. The pipeline uses resources from Ensembl (17), NCBI (13), the Mouse Genome Database (18) and the UCSC mouse genome database (14). The following describes the workflow of the pipeline.

(i) Mouse genomic sequences flanking RISs from tumors are BLAT searched (7) against the public draft of the mouse genome sequence.

(ii) The best BLAT hit is examined and the chromosome and nucleotide position of each integration site on the mouse genome sequence is determined.

(iii) CISs are identified using current CIS definitions (9).

(iv) For retroviral integrations not located at CISs, data from the NCBI are queried in order to identify the gene(s) located nearest the integration site that is the best disease candidate.

(v) For retroviral integrations located at CISs, data from NCBI and Ensembl are queried in order to identify the gene or genes located closest to the CIS. The best candidate is then picked based on the nature of its gene product (i.e. is it a good disease gene candidate?) and the orientation/position of the retroviral integrations relative to the candidate gene. Retroviral integrations are generally, but not universally,

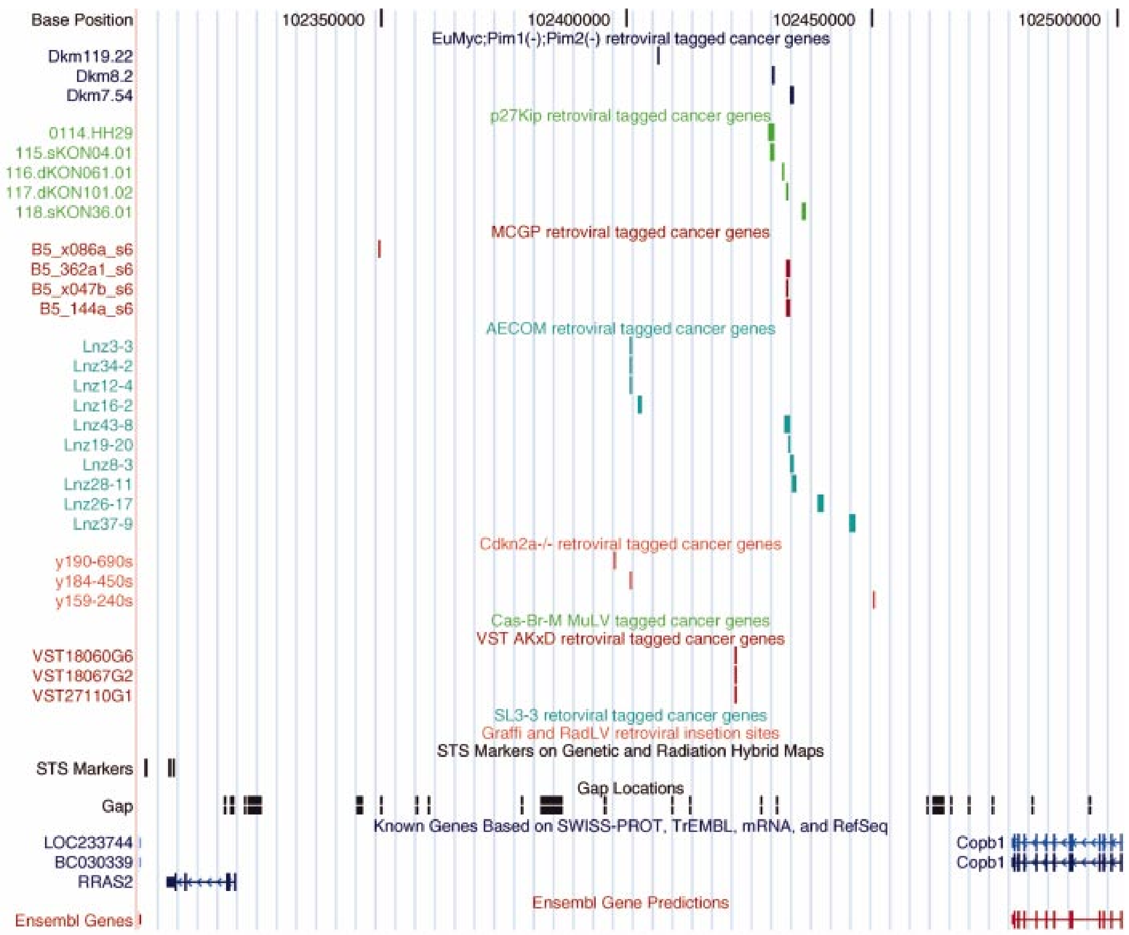

Base Position
102350000
102400000
102450000
102500000
EuMyc;Pim1(-);Pim2(-) retroviral tagged cancer genes
Dkm119.22
Dkm8.2
Dkm7.54
p27Kip retroviral tagged cancer genes
0114.HH29
115.sKON04.01
116.dKON061.01
117.dKON101.02
118.sKON36.01
MCGP retroviral tagged cancer genes
B5_x086a_s6
B5_362a1_s6
B5_x047b_s6
B5_144a_s6
AECOM retroviral tagged cancer genes
Lnz3-3
Lnz34-2
Lnz12-4
Lnz16-2
Lnz43-8
Lnz19-20
Lnz8-3
Lnz28-11
Lnz26-17
Lnz37-9
Cdkn2a-/- retroviral tagged cancer genes
y190-690s
y184-450s
y159-240s
Cas-Br-M MuLV tagged cancer genes
VST AKxD retroviral tagged cancer genes
VST18060G6
VST18067G2
VST27110G1
SL3-3 retorviral tagged cancer genes
Graffi and RadLV retroviral insetion sites
STS Markers on Genetic and Radiation Hybrid Maps
STS Markers
Gap Locations
Gap
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
LOC233744
BC030339
RRAS2
Copb1
Copb1
Ensembl Gene Predictions
Ensembl Genes

Figure 3. Screenshot of the UCSC Mouse Genome Browser Version mm3 (http://genome.ucsc.edu) with RTCGD custom annotation tracks. Viral integration sites from nine groups using various different mouse tumor models are labeled by different colors and users can visually check the candidate genes for each CIS.

oriented in the reverse orientation relative to a disease gene when they are located upstream from the gene and in the same orientation when they are located downstream from the gene

(vi) The results of this database mining are then deposited into a temporary database where they are manually reviewed by researches based on the literature and known information regarding gene function and structure. Data that pass this test are deposited in the RTCGD.

The sequence analysis pipeline significantly decreases the time required for annotating RISs and gives researchers more time to focus on the CIS genes themselves.

FUTURE DEVELOPMENTS

During the next year, we plan to identify and release data for more than a thousand new RIS sequences. We also plan to add two new search tools to our website. One is a tool to identify CIS genes that are highly specific to a specified tumor model or tumor type. Another tool will display the pathways in which each CIS gene is known to function (when these data are available). Links are also provided to the KEGG (19) and Gene Ontology (20) databases in order to help facilitate annotation. Several CIS genes have been shown to function in pathways already associated with hematopoietic disease such as the Ras, Notch, Jak/Stat and Nfkb pathways (9). Others, however, appear to function in pathways not yet associated with hematopoietic disease such as the Wnt signaling pathway (9). While the Wnt signaling pathway has been associated with other human cancers, including colorectal cancer, desmoid tumors and hepatoblastomas (21), it has not yet been causally associated with human hematopoietic disease. This search tool should make it possible to identify additional pathways not yet associated with hematopoietic disease. Knowledge of such pathways will ultimately allow the development of better treatments for these human diseases.

ACKNOWLEDGEMENTS

We thank the National Cancer Institute at Frederick, Advanced Biomedical Computing Center for hosting and supporting RTCGD. This research was supported by the DHHS, NCI.

REFERENCES

1. Mikkers,H. and Berns,A. (2003) Retroviral insertional mutagenesis: tagging cancer pathways. *Adv. Cancer Res.*, **88**, 53–99.

2. Li,J., Shen,H., Himmel,K.L., Dupuy,A.J., Largaespada,D.A., Nakamura,T., Shaughnessy,J.D.,Jr, Jenkins,N.A. and Copeland,N.G. (1999) Leukaemia disease genes: large-scale cloning and pathway predictions. *Nature Genet.*, **23**, 348–353.
3. Silver,J. and Keerikatte,V. (1989) Novel use of polymerase chain reaction to amplify cellular DNA adjacent to an integrated provirus. *J. Virol.*, **63**, 1924–1928.
4. Sorensen,A.B., Duch,M., Jorgensen,P. and Pedersen,F.S. (1993) Amplification and sequence analysis of DNA flanking integrated proviruses by a simple two-step polymerase chain reaction method. *J. Virol.*, **67**, 7118–7124.
5. Valk,P.J., Joosten,M., Vankan,Y., Lowenberg,B. and Delwel,R. (1997) A rapid RT-PCR based method to isolate complementary DNA fragments flanking retrovirus integration sites. *Nucleic Acids Res.*, **25**, 4419–4421.
6. Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
7. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
8. Hwang,H.C., Martins,C.P., Bronkhorst,Y., Randel,E., Berns,A., Fero,M. and Clurman,B.E. (2002) Identification of oncogenes collaborating with $p27^{Kip1}$ loss by insertional mutagenesis and high-throughput insertion site analysis. *Proc. Natl Acad. Sci. USA*, **99**, 11293–11298.
9. Suzuki,T., Shen,H., Akagi,K., Morse,H.C., Malley,J.D., Naiman,D.Q., Jenkins,N.A. and Copeland,N.G. (2002) New genes involved in cancer identified by retroviral tagging. *Nature Genet.*, **32**, 166–174.
10. Kim,R., Trubetskoy,A., Suzuki,T., Jenkins,N.A., Copeland,N.G. and Lenz,J. (2003) Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. *J. Virol.*, **77**, 2056–2062.
11. Lund,A.H., Turner,G., Trubetskoy,A., Verhoeven,E., Wientjens,E., Hulsman,D., Russell,R., DePinho,R.A., Lenz,J. and van Lohuizen,M. (2002) Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. *Nature Genet.*, **32**, 160–165.
12. Kroon,E., Krosl,J., Thorsteinsdottir,U., Baban,S., Buchberg,A.M. and Sauvageau,G. (1998) Hoxa9 transforms primary bone marrow cells through specific collaboration with Meis1a but not Pbx1b. *EMBO J.*, **17**, 3714–3725.
13. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
14. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
15. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
16. Baldarelli,R.M., Hill,D.P., Blake,J.A., Adachi,J., Furuno,M., Bradt,D., Corbani,L.E., Cousins,S., Frazer,K.S., Qi,D. *et al.* (2003) Connecting sequence and biology in the laboratory mouse. *Genome Res.*, **13**, 1505–1519.
17. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
18. Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A. and Eppig,J.T.; Mouse Genome Database Group. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.
19. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
20. Yeh,I., Karp,P.D., Noy,N.F. and Altman,R.B. (2003) Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics*, **19**, 241–248.
21. Taipale,J. and Beachy,P.A. (2001) The Hedgehog and Wnt signalling pathways in cancer. *Nature*, **411**, 349–354.